Talker Differences and Gender Effects in Audio-Visual Speech Perception

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation *with research distinction* in Speech and Hearing Science in the undergraduate colleges of the The Ohio State University

By

Amanda Hager

The Ohio State University

April 2013

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Listeners integrate auditory and visual cues in perception of speech when communicating in both normal and compromised listening environments. Three factors affect the success of interaction in communication situations: characteristics of the talker, characteristics of the listener, and characteristics of the speech signal itself. In everyday life, individuals must comprehend speech produced by many different talkers. Little, however, is known about the characteristics of talkers that make them more intelligible and that best facilitate audio- visual integration. In the present study, 10 adult listeners, with normal or corrected-to-normal vision and auditory thresholds at or better than 25 dB HL across all frequencies, were presented with everyday sentences produced by eight different talkers selected from a commercially available software package (HeLPs, Sensimetrics, Inc.). Sentences were presented under audio-only, visual-only, and audio + visual modalities. Talkers varied widely in gender, age, and ethnicity. Auditory input was degraded to approximate a sloping hearing loss (55 dB HL at 1000 Hz). Results showed significant differences across talkers, but no males were more intelligible in auditory-only presentation, whereas females were more intelligible under visual-only and audio+visual presentation. These results provide new insights for the design of oral rehabilitation programs for hearing-impaired persons.

Acknowledgements

I would like to thank my research advisor, Dr. Janet Weisenberger. Through her patience and guidance, my thesis experience has been an exciting and fulfilling endeavor. I would not have had such a satisfying experience without her as a mentor. Her assistance and mentorship is extremely appreciated. I would also like to thank my subjects for their time, optimism, effort, and cooperation.

This project was supported by an SBS Undergraduate Research grant.

Table of Contents

Chapter 1: Introduction and Literature Review

Although speech perception is commonly recognized as an auditory-only signal, in reality it is a multimodal process that incorporates both auditory and visual input. In situations where the auditory signal is compromised, such as either a noisy environment or a hearing impaired individual, the perception of speech can be very difficult and may require the use of both auditory and visual modalities. The visual stimuli presented in this integration can greatly increase the intelligibility of the speech signal. The McGurk and MacDonald (1976) study, however, suggests that information from visual stimuli is integrated with the auditory stimuli even when the speech signal is perfectly intelligible. This integration, known as the *McGurk effect*, is present even after observers are aware of the concept. Listeners in their study would fuse the visual + audition place of articulation of two different sounds even with a completely intelligible auditory signal. For example, an auditory signal like the bilabial /ba/ would be dubbed onto the video image of a talker saying velar sound /ga/. The listener would integrate this information and perceive the sound as /da/. Results of this study validate that audiovisual integration is constantly occurring in speech perception.

Grant and Braida (1991) observed that the addition of a visual speech signal to a degraded auditory signal can improve speech intelligibility in hearing-impaired individuals. In order to determine the factors influencing the widely differing audio-visual integration abilities across persons, Grant and Seitz (1998) examined separately in hearing-impaired individuals the improvement produced for consonants, words, and sentences in “congruent” and “discrepant” auditory-visual nonsense syllables. In

congruent speech, the acoustic speech signal matches the presented visual speech signal. Conversely, discrepant speech is demonstrated when an auditory speech signal is dubbed onto a visual signal of a different sound. Grant and Seitz found that the intelligibility of meaningful sentences benefits more from the presentation of a visual signal than the identification of nonsense syllables (content-free stimuli). Both visual place cues and auditory manner cues were determining factors in audio+visual overall recognition, and the ability to integrate cues from both of these modalities varied greatly across participants. This study also showed that the benefit visual cues have on the speech recognition of audio+visual integration can be interpreted in terms of degree of redundancy of auditory and visual cues, with greater redundancy related to smaller benefits. These findings suggest that performance in audio-visual integration could not be predicted based on unimodal processes such as auditory and visual-only performance, leading Grant and Seitz to argue that integration is a process independent of auditory-only and visual-only processing.

In his three experiments, Brancazio (2004) observed influences of lexical information in audiovisual speech perception, testing 62 undergraduate students with both words and non-words, varying in initial consonant and place of articulation. Results indicated that the McGurk effect (a visual signal contribution), noted above, was more prevalent when it was formed in a word, than in a non-word. For the responses listeners provided, lexical effect was larger for slower than for faster responses. Results support models that suggest that lexical processing of speech is a process independent of modality. (Brancazio, 2004). Their findings support the ideas that the integration of

visual input (McGurk effect) is more prevalent in meaningful sentences and that performance in one modality cannot predict performance in another.

In any communication situation, there are three factors that can affect the success of the interaction: characteristics of the talker, characteristics of the listener, and characteristics of the speech signal itself. Considerable research has investigated the characteristics of the acoustic speech signal, as well as the individual differences in listeners, such as degree of hearing loss, that can affect communication. Far less attention has focused on the characteristics of individual talkers that make some more intelligible than others. Buchan et al. (2008) investigated how gaze behavior varies across talker identity in audiovisual speech perception. The behavioral strategies that are involved in gathering visual information by watching faces in a communication situation were evaluated by using an eyetracker to measure the movement of gaze. Noise was added to the speech signal (simulating a compromised communication situation) and a noticeable effect was seen for location and duration of fixations. Eye gaze became more centralized to the face during this noise presentation and less on the mouth. In listening situations where noise was not added and a different talker was presented every trial, focus of eye gaze stayed more on the mouth, as opposed to using the same talker for consecutive trials. (Buchan et al., 2008). The study helps demonstrate that perception of visual information while communicating becomes more significant when the auditory signal is degraded. This study further indicates the significance of visual characteristics of the talker that are important in multiple ways for different communication situations.

Kim and Davis (2011) also showed of the importance of visual characteristics of talkers in a degraded speech signal where the familiarity of both the voice and face of

talkers was tested in a speech intelligibility task. In this study, 34 participants were trained to recognize both the voice and face of four animated characters, until ceiling levels were reached. After training, a speech-in- noise task was given to the listeners and tested in 3 types of condition: familiar voice and familiar face, familiar voice and unfamiliar face, and unfamiliar voice and face. Results indicated that while speech perception was better in the familiar voice and face condition in comparison to the unfamiliar voice and face condition (a talker familiarity effect), the familiar voice and unfamiliar face scores did not significantly differ from the unfamiliar voice and face conditions. (Kim and Davis, 2011). These results support the idea that the familiar voice effect was reduced because of its pairing with an unfamiliar face, and underscore the importance of varying talker characteristics for both familiar and non-familiar talkers.

Andrews (2007) addressed talker characteristics by investigating talker intelligibility for degraded auditory speech syllables. In her study, ten normal hearing listeners were presented with auditory, visual, and audiovisual speech syllable stimuli produced by 14 talkers. The talker productions varied widely in intelligibility in the degraded auditory-only and audiovisual presentations. However, little variability was observed for visual-only presentations. Andrews also measured the amount of audiovisual integration, defined as the difference in performance between audiovisual presentation and the best single modality, either auditory-only or visual-only. Interestingly, “the talkers producing the most audiovisual integration were not those with the highest auditory-only intelligibility”(Andrews, 2007, p. 2).

One way to maximize audio-visual speech perception is to use a specific speaking style known as “clear speech” when communicating with hearing impaired individuals.

Clear speech is an example of a characteristic of the speech signal itself that can affect a communication interaction. This speaking style differs linguistically and acoustically from conversational speech. Krause and Braida (2002) note that clear speech manifests a slower speaking rate, greater temporal modulation, increased range of voice fundamental frequency, an expanded vowel space, and more stimulus energy in high frequencies. These characteristics have been verified empirically (Chen, 1980; Picheny et al., 1985; Uchanski et al., 1992; Payton et al., 1994.).

Differences in intelligibility across talkers can also affect the perception of clear speech. Gagné et al. (1994) argue that "some individuals do not inherently know how to produce clear speech (or worse, their attempts to produce clear speech resulted in poorer speech perception scores in one or more sensory modality)"(Gagné, 1994, p. 155). Their study was conducted using 10 female talkers who were told to "articulate each word clearly as if you were talking to someone who had difficulty understanding what you are saying"(Gagné, 1994, p. 155). These speech utterances were presented to normal-hearing listeners under degraded auditory, visual, and audio-visual conditions using a hearing-loss simulator (HELOS: Gagné and Erber, 1987) to simulate a sensorineural hearing loss. Results indicated huge variability in intelligibility across talkers. Some of the talkers did not produce a more intelligible clear speech effect in any sensory modality, while others were successful in one but not all modalities.

Ferguson et al. (2004) noted that Gagne et al. had presented speech from female talkers only. Ferguson et al. also investigated the intelligibility of clear and conversational speech, but included 41 talkers of both genders. All talkers were organized into four age brackets: (1) 18 to 24 years, (2) 25 to 31 years, (3) 32 to 38 years, and (4) 39

to 45 years. It was hypothesized that older talkers would have more effective "clear speech" because of potential frequent contact with hearing impaired individuals. Ferguson et al. concentrated on the acoustic speech signal, using degraded auditory-only presentation rather than visual or audiovisual presentations. Results showed that while age and experience speaking to hearing impaired individuals had no effect on talker intelligibility in clear speech, gender did. Female talkers had greater overall (78% versus 68%) vowel intelligibility than males.

Because Ferguson et al. did not examine visual characteristics of talkers, the question remains whether these gender advantages would be prevalent in visual-only and audio-visual presentation as well as in auditory-only speech. The present study addresses this question by investigating differences in talker intelligibility across male and female talkers in three different modalities: degraded audio, visual, and degraded audio +visual. Utterances degraded to simulate a sloping 55 dB high frequency hearing impairment at 1000 Hz were produced using the software program HeLPs, A hearing loss and prosthesis simulator, created by Sensimetrics Corporation. Eight pre-recorded talkers, (4 male and 4 female) were presented, speaking 20 sentences each. Sentences were presented to 8 normal hearing listeners in these three modalities. Results were evaluated to determine if talkers who produce the greatest intelligibility in auditory speech perception also produce similar results in visual and audio-visual presentation. In addition, the question of whether high auditory-only intelligibility or high visual-only intelligibility produces the largest degree of intelligibility in audio-visual integration was assessed.

From past studies, several predictions can be made regarding the possible outcomes of the present study of talker characteristics. The findings of Ferguson et al.

(2004) support the prediction of gender effect, wherein female talkers will have better overall intelligibility than males. The results of Grant and Seitz (1998) and Andrews (2007) support the hypothesis that high auditory-only intelligibility and high visual-only intelligibility are unrelated, as well as high auditory intelligibility having no effect on audio-visual integration. The present investigation should yield additional insights into the factors that make speech most intelligible to hearing impaired persons.

Chapter 2: Method

Participants

Before participants were recruited, this study was approved by The Ohio State University, Institutional Review Board, IRB Protocol number 2012B00045. Participants in this study included 8 listeners who shown normal hearing thresholds of 25 dB HL or better and reported normal or properly aided vision. All were native speakers of American English. All participants received $10 an hour for their participation in this study.

Stimulus Presentation

HeLPs

The stimuli used in this study consisted of 368 randomized sentences produced by prerecorded talkers in the HeLPs software program. HeLPs –Hearing Loss and Prosthesis Simulator (Sensimetrics Corporation), is a computer software program that simulates the auditory communication difficulties associated with hearing loss and the potential benefits provided by hearing aids and cochlear implants. This new product permits simulation of any degree or pattern of hearing loss, differentiation between air and bone conduction for left or right ears, as well as simulation of tonal or noisy tinnitus. HeLPs provides a graphic interface on the computer for controlling the simulation and a set of calibrated headphones for listening to the simulator's output. This program provides presentations in three modalities: audio only, visual only, and audio+visual processed by the simulator in order to allow the listener to use reading speech. The characteristics of hearing and prosthesis are specified separately for the left and right

sides, with controls for selecting loss and prosthesis settings, talkers, background noises, and reverberation.

Talkers

Within HeLPs, there are 10 recorded talkers (5 male, 5 female), varying widely in ethnicity with ranges of 13 to 67 sentences per talker available. These generic sentences consist of statements and questions, and are designed around everyday conversation topics such as places, people, and events. In this study we utilized 8 talkers (4 male and 4 female), presenting 20 randomized sentences for each.

Presentation

These recorded sentences were produced to every listener in 3 different modalities: auditory only, visual only, and audio+visual. These three presentations are available within the software by clicking their respective option on the screen before playing the talker's sentence.

Auditory-only Presentation

Sentences were presented with an auditory stimulus and no visual stimulus. We selected a hearing loss configuration provided by the software that simulated a sloping 55 dB high frequency hearing loss at 1000 Hz, presented at 70 dB through Sennheiser supra-aural headphones.

Visual-only Presentation

The visual-only condition consisted of a 5x4 in. video image presentation of the talker delivering the sentence on a computer monitor with no auditory stimulus provided.

Audio+Visual Presentation

Audio+visual presentation consisted of both audio and visual options as described above.

Headphones Calibration

Headphones used for stimulus presentation were calibrated acoustically using a KEMAR manikin. By using continuous speech shaped noise at 65 dB SPL, with anechoic settings, a steady noise stimulus (with brief occasional pauses) was presented and measured by the equipment.

Test Enclosure

Testing for this study took place in a quiet lab room in the basement of Pressey Hall. Listeners were tested on weekends and evenings to insure a consistently quiet environment.

Procedure

After obtaining informed consent was obtained from the participants, they were seated in a quiet lab room in the basement of Pressey Hall. The experimenter provided with written and spoken instructions for the procedure. Both the participant and the

experimenter then put on headphones from the same laptop, using a dual headphone jack splitter and the participant was provided a demonstration of the software program (HeLPs).

After the demonstration was completed, listeners, based on random assignment, were then presented with either a visual or audio-only version of a sentence, followed by audio-only or visual only. An audio+visual presentation of the same sentence was always the third modality the listeners heard. Up to 3 repetitions were allotted for each sentence presented under each modality. After each presentation, listeners were asked to replicate what they perceived as the content of the sentence. The examiner then wrote down the listener responses for each condition on a score sheet. No feedback was given to the participants at any point on the accuracy of their responses. A total of 20 sentences were presented in this way for each of the 8 talkers.

The order in which the talkers were presented was based on random assignment across participants, as was the determination of whether auditory only or visual only occurred first for each sentence. For a given talker, the selection of a sentence for each trial was controlled by the software using a randomized with replacement algorithm.

Frequent rest breaks were encouraged for the participants to reduce fatigue. No more than 1-2 hours a day were spent per listener. Total testing time for each listener was approximately 6 hours.

Chapter 3: Results and Discussion

Results were analyzed in order to determine if there was an overall gender effect between females and males in intelligibility to hearing impaired persons. Figures 1 and 2 show the results averaged across all listeners for the 4 female talkers and 4 male talkers under the audio-only, visual-only, and audio+visual modalities. Results indicate that males yielded slightly higher audio-only intelligibility while females yielded higher visual-only and audio+visual intelligibility scores, as shown in Figure 3.

In order to determine whether there was an overall gender effect, a 2 factor repeated measures analysis of variance (ANOVA) with arcsine transformed data was performed for the factors of gender and modality. Results indicated no significant main effect of gender, $F(1,7) = .004$, although a trend was present. However, a significant effect could be found in modality, $F(2,14) = 245.287$, $p < .001$, wherein audio+visual conditions yielded the highest scores and audio-only conditions, the lowest. Also, a significant interaction between gender and modality was found, $F(2,14) = 4.80$, $p = .026$, wherein female talkers were more intelligible in visual-only and audio+visual modalities and male talkers were more intelligible in the audio-only modality. These results are interestingly different from the findings of Ferguson et al. (2004); female talkers in her study produced significantly higher scores in an audio-only condition, whereas in the present study males had higher audio-only scores and females had higher visual-only and audio+visual scores.

To get a better sense of differences across talkers that contribute to overall findings, the data for each of the talkers were analyzed. Figures 1 and 2 indicate that Joe had the highest intelligibility scores in the auditory-only modality, and Lorraine had the

highest intelligibility scores in the visual-only and audio+visual modalities. A 2 factor repeated measures ANOVA with talker and modality as factors indicated significant differences across talkers, $F(7,49) = 9.015$, $p<.001$, and a significant effect across modality, $F(2,14) = 216.541$, $p<.001$. Also, a significant interaction was found between talkers and modality, $F(14, 98) = 5.394$, $p<.001$.

These results are indicative of the factors that make for better intelligibility in different modalities for individual talkers. One characteristic of clear speech noted above is an overall slower speech rate (Uchanski et al., 1992), and in the present study the slowest speech rate was exhibited by Joe with an average sentence time of 3.2 seconds. In contrast, the fastest speaking rate belonged to Alan (average of 1.5 seconds), whose intelligibility was poorest in every modality. This could be one possible explanation for why Joe was a highly intelligible talker, and Alan was not. A second possible explanation for differences across talkers might be the degree of experience in speaking to hearing impaired individuals. Lorraine is a researcher who has worked extensively with hearing impaired listeners thus has a long experience in communication with them. As noted, Lorraine was highly intelligible under all 3 modalities. The degree of experience in speaking to hearing impaired listeners for the other talkers is not known. Ferguson et al. (2004) also investigated experience talking to hearing impaired persons. In her analysis she equated experience to the talkers' age rather than known experience, and found no significant effect. In contrast, in the present study the oldest talker was Alan, who was also the least intelligible talker. This suggests that talker age may not equate with experience communicating with hearing impaired persons.

Table 1 shows that the amount of audio-visual integration (difference between audio+visual and best single modality) differed across talkers wherein Ann and Christina exhibited the most, and Alan, Angelo, and Samantha yielded the least, integration. Talkers with the highest audio+visual scores were not necessarily the ones who produced the highest amount of audiovisual integration. The reason may be because the visual-only performance was so high for the two most intelligible talkers, Lorraine and Joe. Figure 1 also shows that there is only one talker, Bob, who had higher audio-only intelligibility scores than visual-only scores. Subjects complained that he was difficult to lip read because of his large lips in the visual-only condition. Christina, who was also a less intelligible talker, also received similar complaints about the size of her lips.

Other potential insights into characteristics that make talkers more intelligible may come from anecdotal comments made by listeners during testing. Joe, one of the most intelligible talkers, had little facial expression, spoke very slowly, and was described as “boring”. Listeners noted that Alan, the least intelligible talker, did not move his lips enough and some found a sideways lip motion displayed that was very distracting. He presented with a pronounced Bostonian accent that was most notable in some vowel pronunciations, and was overall the most disliked talker. Angelo was observed to have a very expressive face, but was noted to appear to be speaking very rapidly in the visual-only condition and this made him difficult to lip-read. Samantha was reported by some of the subjects to be so expressive in her face that they found it hard to concentrate on the utterance.

Chapter 4: Summary and Conclusion

Overall, the results of the present study indicate that while a significant gender effect wherein females are more intelligible than males was not found, effects across talkers were. There were significant differences across talkers in intelligibility scores; some talkers seemed to be more intelligible based on temporal aspects of the speech signal similar to clear speech, and expressive facial movements during utterances. A significant effect across modalities was found, in which the audio+visual modality produced the highest intelligibility, and the audio-only condition, the lowest. A significant interaction was found between talkers and modality; Lorraine yielded the highest visual-only and audio+visual modality intelligibility scores, and Joe produced the highest audio-only modality intelligibility scores. A significant interaction was also found between gender and modality, wherein male talkers were more intelligible in the audio-only condition, and females in the visual-only and audio+visual conditions. A variety of possible explanations can be evaluated as to the levels of intelligibility of all the talkers.

There were a number of challenges in using the HeLPs software program for the present study. Each of the 8 talkers used in the testing, produced a different set of sentences. This did not allow us to use the same sentence set across all talkers to present to our subjects. Although not quantified, it appeared not all sentence sets were of equal difficulty. Some of the sets seemed to be thematically arranged, and this could have facilitated high intelligibility. For example, Lorraine's sentence set contained multiple sentences about geography, inquiring several times about the capital of certain states and countries.

The present study tested only 8 listeners because of time constraints, but future work might expand the number of listeners exposed to each talker and perform more quantitative analysis of acoustic and visual characteristics of sentences spoken by each talker. Acoustic analysis could look at the vowel formant values and consonant-vowel formant transitions, visual spatial frequency analysis, as well as measure of the degree of lip opening, could also provide insights.

Studies of talker differences can have important implications for the design of aural rehabilitation training programs. The existence of cross-talker differences is very important and should be a feature in any computerized training. It is also important that the listener be exposed to a number of talkers. Knowledge of what makes a talker more intelligible can allow easier talkers to be presented first, and more difficult talkers later to minimize frustration during training.

In an aural rehabilitation situation, family members and significant others of the hearing-impaired person are often given a general set of instructions on how they can speak more intelligibly and be best understood by their loved one. From looking at the results of the present study, certain instructions like decreasing the speaking rate of the speech signal could be included. Also, instructions might be provided about maintaining expressiveness of the entire face while speaking, but without being overly expressive to the point of distraction. Talkers with big lips can be advised that this may make them more difficult to understand and to be very aware of their lip movements and mouth opening. Overall, the present results provide new insights for the design of aural rehabilitation programs for hearing-impaired persons.

References

Andrews, B. (2007). Auditory and visual information facilitating speech integration. Undergraduate honors thesis, The Ohio State University.

Brancazio, L., (2004). Lexical Influences in Audiovisual Speech Perception. *Journal of Experimental Psychology 30,* 445-463

Buchan, J.N., Paré, M., Munhall, K. G., (2008). The Effect of Varying Talker Identity and Listening Conditions on Gaze Behavior During Audiovisual Speech Perception. *National Institute of Health Public Access 1242,* 162-171

Chen, F. (1980). Acoustic characteristics of clear and conversational speech at the segmental level. Cambridge, MA: Massachusetts Institute of Technology.

Ferguson, S. (2004). Talker differences in clear and conversational speech: Vowel Intelligibility for normal-hearing listeners. *J. Acoustic Society of America 116,* 2365-2373.

Gagné, J.P., and Erber, N.P. (1987). Simulation of sensorineural hearing loss. *Ear and Hearing, 8,* 232-243.

Gagné, J.P., Masterson, V., Munhall, K.G., Bilida, N., and Querengesser, C. (1994). Across talker variability in auditory, visual, audiovisual speech intelligibility for conversational and clear speech. *J Academy Rehab 27*, 135-158.

Grant, K.W., and Braida, L.D., (1991). Evaluating the articulation index for audiovisual input. *Journal of the Acoustical Society of America 89*, 2952-2960.

Grant, K.W., and Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America 102,* 2677-2690.

Kim, J., Davis, C., (2011). Testing Audiovisual Familiarity Effects on Speech Perception In Noise.

International Congress of Phonetic Sciences 17, 17-21

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices.

Nature, 264, 746-748.

Payton, M.A., Uchanski, R.M., and Braida, L.D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired impaired hearing. *The Journal of the Acoustical Society of America 95,* 1581-1592.

Picheny, M.A., Durlach, N.I., and Braida, L.D., (1985). Speaking clearly for the hard of hearing; Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research 28,* 96-103.

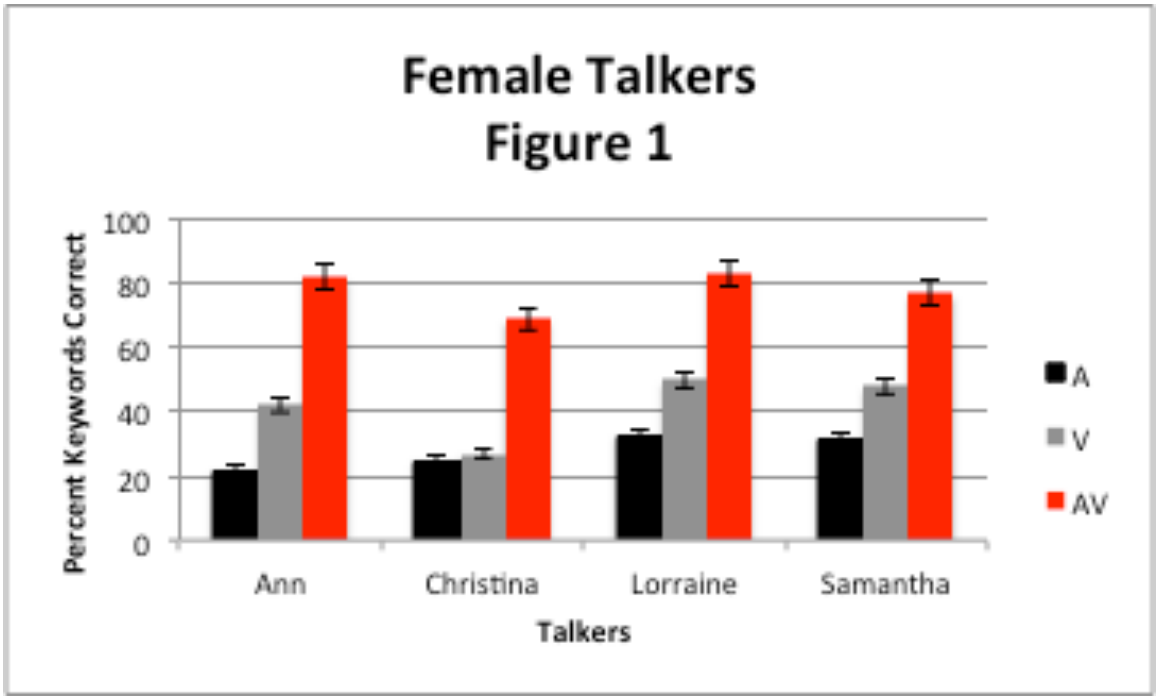
Uchanski, R.M., Miller, K.M., Reed, C.M., and Braida, L.D. (1992). Effects of token variability on vowel identification. In Schouten, E.H. (Ed.), *The auditory processing of speech.* New York: Mouton de Gruyter.

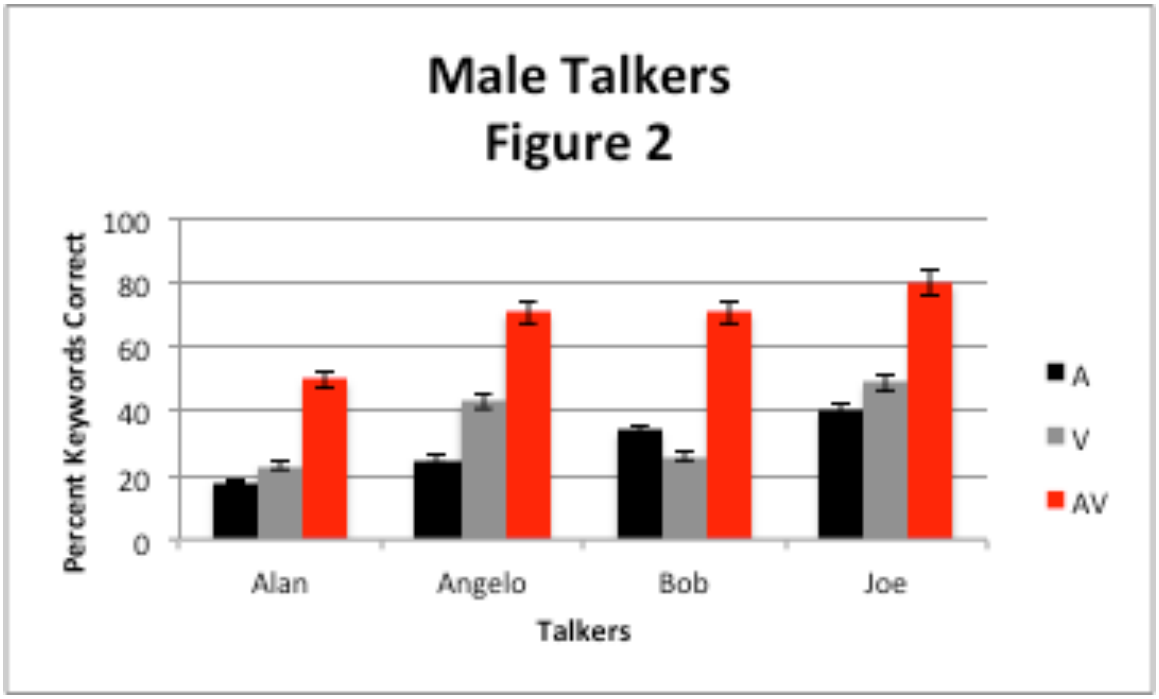
List of Figures/Tables

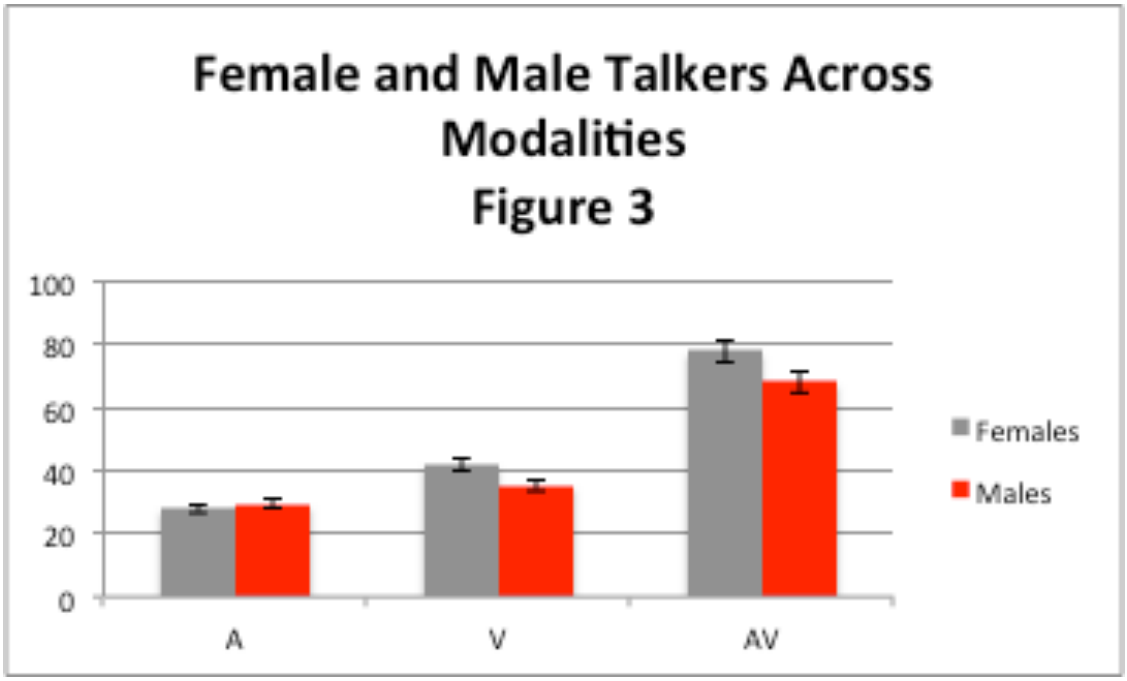
Percent Audiovisual Integration for Female and Male Talkers

Table 1

Ann	40
Christina	42
Lorraine	33
Samantha	29
Alan	27
Angelo	28
Bob	37
Joe	31


Female Talkers
Figure 1
Percent Keywords Correct
100
80
60
40
20
0
Ann
Christina
Lorraine
Samantha
Talkers
A
V
AV


Male Talkers
Figure 2
Percent Keywords Correct
100
80
60
40
20
0
Alan
Angelo
Bob
Joe
Talkers
A
V
AV


Female and Male Talkers Across Modalities
Figure 3
100
80
60
40
20
0
A
V
AV
Females
Males